REPRODUCIBILIDAD EN LAS TAREAS DE EVALUACIÓN

Plan de impulso de las Tecnologías del Lenguaje

Paolo Rosso y Francisco Rangel

Centro de Investigación PRHLT

Universitat Politècnica de València

Diciembre 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital y Red.es, que no comparten necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

ÍNDICE DE FIGURAS

REPRODUCIBILIDAD EN LAS TAREAS DE EVALUACIÓN

En el Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital y Red.es, existe el interés para realizar ejercicios de evaluación reproducibles y orientados a componentes. Este estudio manifiesta el interés de la comunidad científica acerca del tema de la reproducibilidad así como demuestra la organización del laboratorio CENTRE sobre reproducibilidad en los tres foros de evaluación más importantes de Europa (CLEF), EE.UU. (TREC) y Japón (NTCIR).

El objetivo de CENTRE es, dada una tarea, poder reproducir los resultados hechos públicos por los participantes, usando los componentes desarrollados anteriormente. En este sentido, en el futuro CENTRE se plantea abordar más en profundidad la reproducibilidad a nivel de componentes de sistemas.

Este estudio quiere ser también un marco de referencia para poder ayudar a los investigadores en la organización futura de tareas de evaluación. Para facilitar la reproducibilidad de los ejercicios de evaluación es importante que en el momento de organizar una tarea de evaluación el investigador considere:

- Un marco común de evaluación (herramienta de evaluación)
- Una evaluación como servicio (plataforma de evaluación)
- Una herramienta para poder compartir el código

En los siguientes aparatados abordaremos todos estos aspectos.

1 INTRODUCCIÓN

El concepto de reproducibilidad se refiere al grado de concordancia entre los resultados obtenidos para un mismo experimento llevado a cabo por dos o más equipos independientes de investigadores, esto es, llevado a cabo por individuos diferentes, en ubicaciones diferentes y/o con instrumental diferente. La reproducibilidad, por tanto, mide la posibilidad de que un experimento sea reproducido por otros investigadores obteniendo resultados similares.

En el caso particular de la experimentación con datos, donde los resultados muestran el rendimiento de un determinado algoritmo sobre un determinado conjunto de datos y dada una determinada medida de evaluación, se pueden dar los siguientes dos escenarios donde la reproducibilidad se ve comprometida:

ESCENARIO 1 (replicabilidad/repetibilidad): Mismo conjunto de datos

- Un investigador A publica un artículo en el que afirma que:
  - El algoritmo X es mejor que el algoritmo Y en el conjunto de datos D.
- Un investigador B replica el experimento pero encuentra que:
  - El algoritmo X es peor que el algoritmo Y en el mismo conjunto de datos D.

ESCENARIO 2 (reproducibilidad): Diferente conjunto de datos

- Un investigador A publica un artículo en el que afirma que:
  - El algoritmo X es mejor que el algoritmo Y en el conjunto de datos D.
- Un investigador B replica el experimento pero encuentra que:
  - El algoritmo X es peor que el algoritmo Y en un conjunto de datos diferente D'.

Es importante distinguir el concepto de reproducibilidad anterior del concepto de repetibilidad. La repetibilidad consiste en ser capaces de repetir un mismo experimento en las mismas condiciones y obtener los mismos resultados. Es decir, que el mismo equipo, en la misma localización y con los mismos instrumentos, sea capaz de obtener los mismos resultados al repetir un experimento previo. Se podría resumir de la siguiente manera:

- Repetibilidad/Replicabilidad: mismas condiciones -> mismos resultados
- Reproducibilidad: similares condiciones -> resultados comparables

Ambos conceptos son importantes ya que permiten su avance de manera sólida. Sin embargo, en este informe nos centramos en el concepto de reproducibilidad, y por tanto no debemos confundirlo con el de repetibilidad.

2 INTERÉS DE LA COMUNIDAD CIENTÍFICA

El interés de la comunidad científica en la reproducibilidad se materializa en la organización del laboratorio CENTRE[1] [1].

Figura 1.ClEf/Ntcir/Trec REproducibility (fuente CENTRE)

[1] Coordinación del laboratorio: Nicola Ferro (Universidad de Padua), Tetsuya Sakai (Waseda University), Ian Soboroff (NIST) http://www.centre-eval.org/

El laboratorio CENTRE aúna tres grandes foros de evaluación (CLEF/NTCIR/TREC) sitos en tres regiones diferentes del globo (Europa / Estados Unidos / Asia):

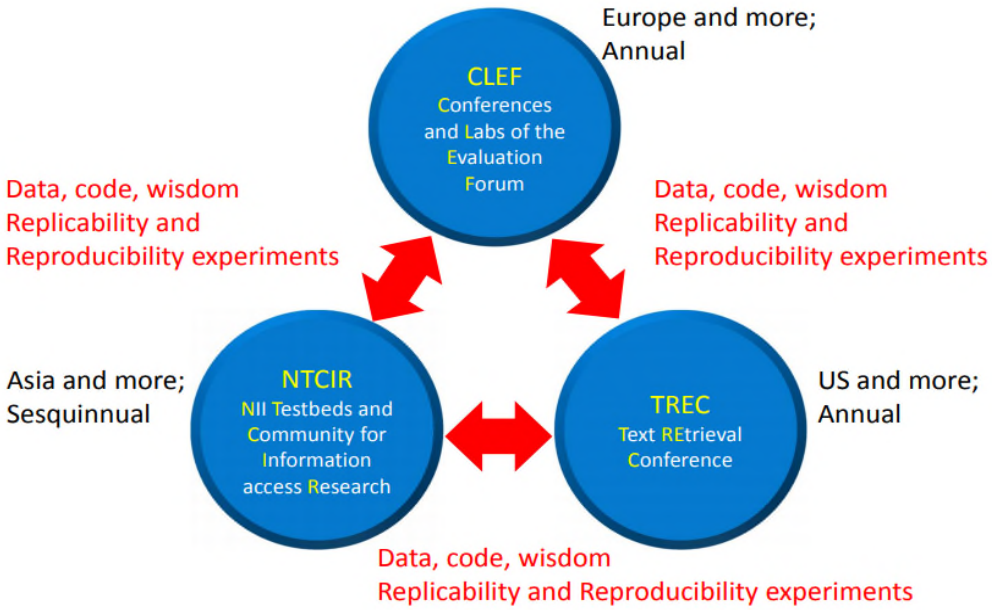

Figura 2. Tarea de reproducibilidad en Europa, EE.UU. y Asia (fuente CENTRE)

El objetivo de CENTRE es lanzar una tarea conjunta en los tres foros de evaluación anteriores persiguiendo:

1. Reproducir los mejores resultados de los mejores (o los más interesantes) sistemas presentados en ediciones previas del CLEF/NTCIR/TREC, usando sistemas de recuperación de información de código abierto estándar.
2. Contribuir a la comunidad con recursos y componentes adicionales desarrollados para reproducir los resultados y con el objetivo adicional de mejorar los sistemas de código abierto existentes.

En particular, CENTRE persigue ahondar en las dos perspectivas de la reproducibilidad mencionadas en la introducción:

- Replicabilidad: diferente equipo, misma configuración experimental.
- Reproducibilidad: diferente equipo, diferente configuración experimental.

Para ello, y como ejemplo, en el caso de CENTRE@CLEF[2] se proponen las siguientes tareas:

[2] http://www.centre-eval.org/clef2018

- Tarea 1 - Replicabilidad: replicabilidad de los métodos seleccionados utilizando las mismas colecciones experimentales.
- Tarea 2 - Reproducibilidad: reproducibilidad de los métodos seleccionados utilizando colecciones experimentales diferentes.
- Tarea 3 - Re-reproducibilidad: usando los componentes desarrollados en las tareas anteriores, replicar/reproducir los resultados hechos públicos por los demás participantes. En este sentido, en el futuro CENTRE se plantea abordar más en profundidad la reproducibilidad a nivel de componentes de sistema.

La reproducibilidad de los experimentos va muy ligada a la metodología de evaluación, y en este sentido es importante disponer de un marco común de evaluación, donde las medidas estén unificadas, así como la posibilidad de proporcionar una infraestructura de evaluación como servicio.

3 MARCO COMÚN DE EVALUACIÓN

La definición de un marco común de evaluación es clave para una correcta valoración de la reproducibilidad de los experimentos. La selección de la medida de evaluación puede dificultar la medición de dicha reproducibilidad. Por ejemplo, el sistema A es mejor que el sistema B en el conjunto de datos D con la medida de evaluación E, mientras que el sistema A es peor que el sistema B en el conjunto de datos D pero con la medida de evaluación E'. La medida E' no tiene por qué ser diferente a la medida E (e.g., *accuracy* vs. *precision*), sino que incluso puede ser una implementación diferente de la misma medida.

En este sentido, herramientas como EVALL[3] [2] proporcionan este marco común de evaluación, permitiendo evaluar un sistema contra colecciones de test previamente almacenadas en la herramienta, sobre un *gold standard* propio o comparando con otro sistema proporcionado. Véase también los entregables desarrollados en el marco del Plan de Impulso de las Tecnologías del Lenguaje [3], [4], [5], [6], [7], [8].

[3] http://evall.uned.es

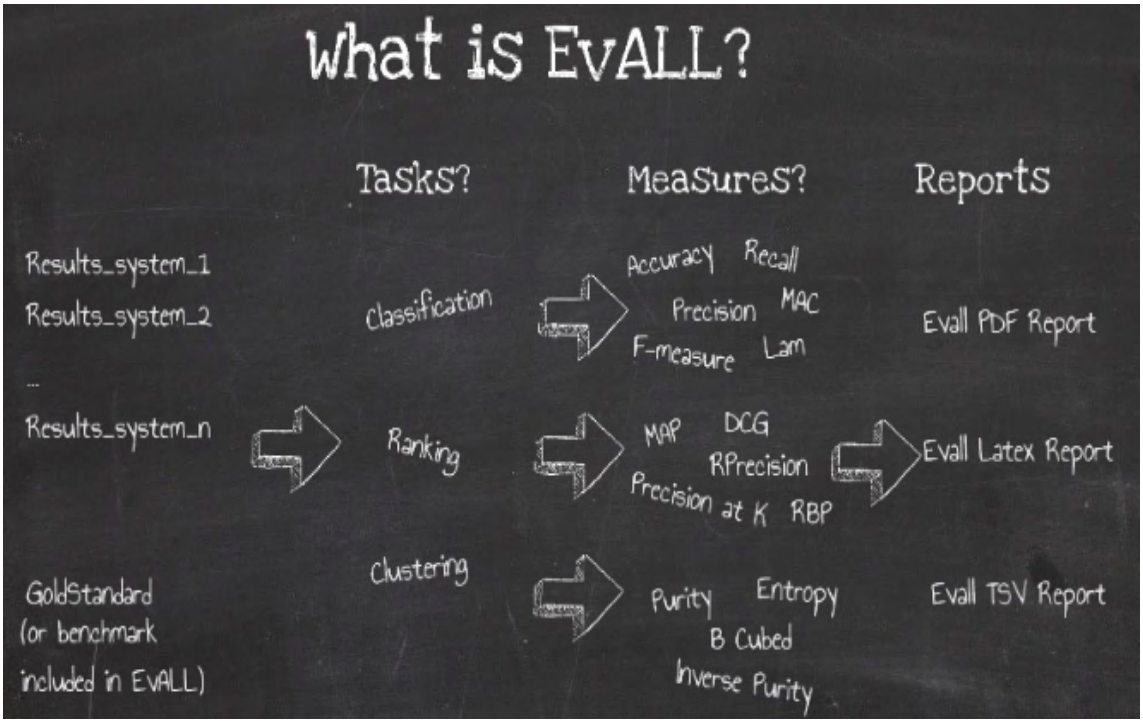

Figura 3. Herramienta de evaluación (fuente: EvALL)

4 EVALUACIÓN COMO SERVICIO

Una tarea compartida (*shared task*) emerge cuando al menos dos sistemas abordan el mismo problema. Sin embargo, el término *shared task* en ciencias de la computación se refiere originalmente a eventos que invitan a investigadores a trabajar en un problema específico, la tarea. Los objetivos para la organización de este tipo de tareas se pueden resumir en:

- Desarrollar nuevas teorías o aproximaciones.
- Implementar softwares adecuados para la tarea.
- Evaluar el rendimiento alcanzable.

Las *shared tasks* se pueden clasificar en tres tipos según lo que se deba enviar (*submission type*): ejecuciones, software, o evaluación como servicio (EaaS). En la siguiente imagen se muestra, para cada uno de estos tres tipos, donde se encuentra la responsabilidad con respecto a los datos y al software, en los participantes, en los organizadores, o en ambos.

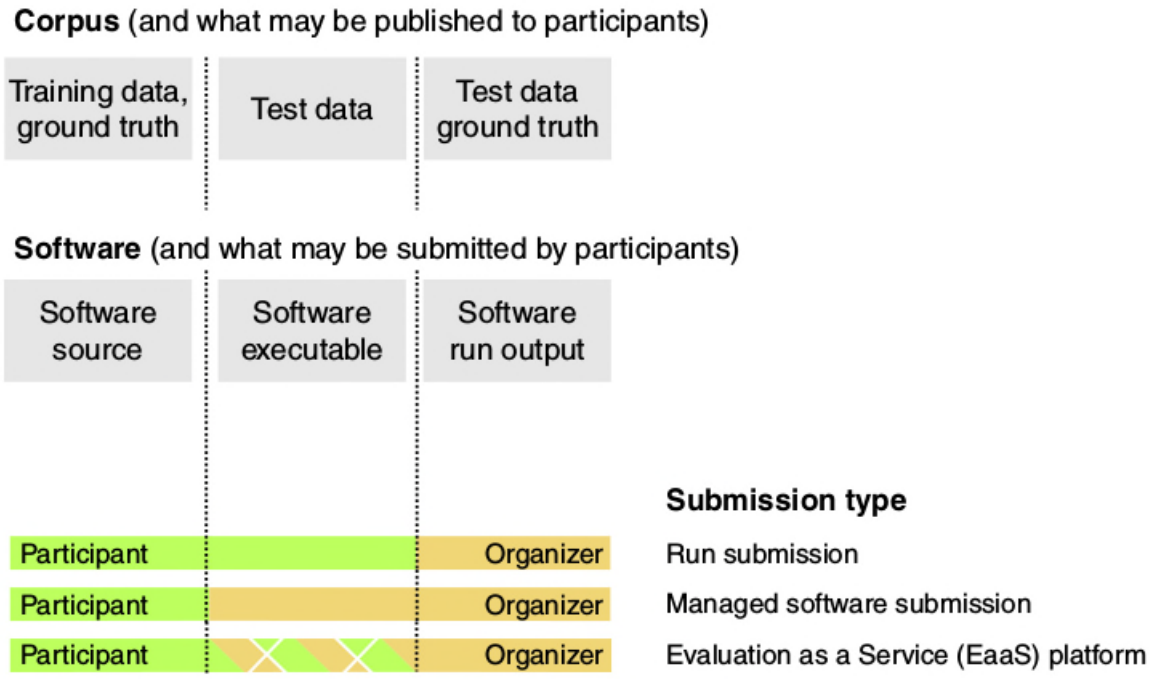

Figura 4. Plataforma de evaluación (fuente: TIRA)

Como se puede apreciar, en el caso de EaaS, la responsabilidad del conjunto de datos de evaluación así como del software ejecutable es compartida entre los participantes y los organizadores. En este sentido la EaaS se enfrenta a una serie de desafíos organizativos, tal y como se muestra a continuación, así como a distintas posibilidades para abordarlos:

1. Diversidad del entorno -> Virtualización: permite soportar una amplia variedad de sistemas operativos y de lenguajes de programación.
2. Ejecución de software no confiable -> Virtualización: permite aislar el sistema a la hora de ejecutar binarios de terceros.
3. Fuga de datos -> Sandboxing: previene la fuga de datos ejecutando el software en un entorno seguro (*securised*).
4. Manejo de errores -> Interfaz de usuario y testeo unitario: proporciona a los usuarios la capacidad de depurar y corregir sus errores.
5. Responsabilidad -> Interfaz de usuario: responsabiliza a los usuarios de su software.
6. Coste de ejecución -> Pago por uso: puede ser a cargo de los organizadores, los participantes o terceras partes.
7. Legal -> Acuerdos de No Divulgación (Non Disclosure Agreement - NDA) individuales o Acuerdos de Licencia de Usuario Final (End User License Agreement - EULA) adecuados. Además, en la actualidad debe adecuarse a la nueva Regulación General de Protección de Datos (RGPD) [9].

En la siguiente tabla se muestran algunas iniciativas que proporcionan EaaS, de las cuales cabe destacar PAN[4] y CoNLL[5] que desde el 2014 y 2015 respectivamente proporcionan EaaS con automatización completa y una interfaz de usuario Web que facilita la interacción de participantes y organizadores con el sistema.

Initiative	Software	Data	Data Access	Submis-sion	Continu-ous	Au-toma-tion	Result Interaction	Technical Support
TREC Microblog 2013-2014	Twitter Tools	Static	API	Result file upload	Fixed deadline	Little	None	Online forum
BioASQ	Dedicated	Static / Dynamic	Download	Result file upload	Fixed deadline	Medium	Online leaderboard	Online forum
VISCERAL Anatomy1/2	VISCERAL Registration System	Static	VM	VM	Fixed deadline	Little	None	Mailing list
VISCERAL Anatomy3	VISCERAL Registration System	Static	VM	VM	Continu-ous	Full	Online leaderboard	Mailing list
CLEF Newsreel	Open Recommendation Platform	Dynamic	API	API	Fixed deadline	Medium	None	Tutorials
CLEF LL4IR	Living Labs API	Static	API / Download	API / Upload	Fixed deadline	Medium	None	Mailing list
C-BIBOP	Codalab	Static	???	???	???	???	???	???
PAN Evaluation Lab	TIRA	Static	VM	VM	Fixed deadline	Full	Web front end	Mailing list
CoNLL Shared Task 2015	TIRA	Static	VM	VM	Fixed deadline	Full	Web front end	Mailing list
TREC Total Recall Track	Baseline Model Implementation	Static	API / Download	VM / Script	Fixed deadline	???	None	Online forum
MIREX	MIREX submission system	Static	None	Compiled Code	Fixed deadline	Little	None	Mailing List

Figura 5. Ejemplo de uso (fuente: TIRA)

En el caso de las tareas mencionadas, ambas cuentan con la plataforma de experimentación TIRA[6] (TIRA Integrated Research Architecture) [10], que responde a la arquitectura mostrada en la siguiente imagen. Son dos los agentes que interactúan con el sistema: los participantes y los organizadores. Por parte de los organizadores, estos son responsables de configurar la tarea (e.g., proporcionar el corpus de evaluación con el gold standard), de la revisión de la tarea y la supervisión del trabajo de los participantes, así como de revisar las ejecuciones de los participantes y publicar sus resultados. Los participantes por su parte se encargan de subir su software y ejecutarlo contra los datasets proporcionados por la plataforma, obteniendo con ellos unas predicciones y su evaluación, y que serán revisadas por los organizadores. TIRA proporciona un sistema de máquinas virtuales con diferentes sistemas operativos (generalmente Windows y Linux), accesibles por diferentes medios (e.g., VNC, ssh), y que dan acceso a los datasets disponibles y permiten a los participantes subir su software, ejecutarlo y obtener resultados. Es decir, los usuarios de TIRA pueden ejecutar y evaluar su software contra cualquier dataset disponible en la plataforma.

[4] https://pan.webis.de
[5] http://www.conll.org
[6] https://tira.io

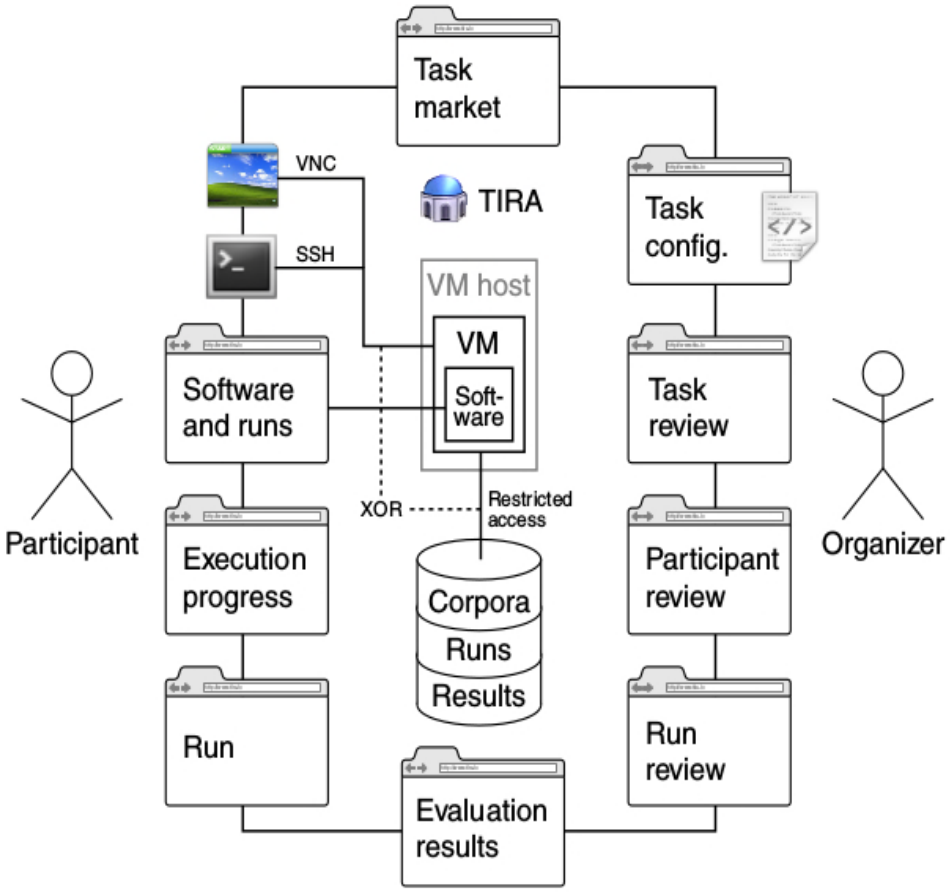

Figura 6. La plataforma de evaluación TIRA (fuente: TIRA)

5 COMPARTICIÓN DEL CÓDIGO

Uno de los principales problemas de la reproducibilidad es que la configuración experimental puede ser muy compleja. En el caso de que la experimentación implique el desarrollo de código, esto puede significar tener que replicar cientos o miles de líneas de programación en la mayoría de los casos complejas.

Sin embargo, por la idiosincrasia propia del software, existen herramientas que permiten compartir estas líneas de código para que otros investigadores las utilicen y se facilite de este modo la reproducibilidad de los experimentos. En este sentido, cada día más investigadores realizan esta compartición de código vía repositorios como GitHub[7], GitLab[8], BitBucket[9], SourceForge[10] o CRAN[11],

[7] https://github.com
[8] https://about.gitlab.com
[9] https://bitbucket.org

entre otros. Además, en ciertos foros, conferencias y laboratorios se promueve la compartición de código, como por ejemplo en el laboratorio PAN que proporciona un repositorio privado[12] a sus participantes para que puedan subir y organizar su código fuente.

6 CONCLUSIONES

Un aspecto que cada vez preocupa más a científicos e investigadores es el de la reproducibilidad. La frontera de la ciencia se estira a partir de la mejora de los métodos existentes y para ello se debe poder reproducir el trabajo previo realizado por otros investigadores. Para asegurar la reproducibilidad es imprescindible disponer de un marco común de evaluación que permita la comparabilidad sin matices. En este sentido existen plataformas que permiten la evaluación de un determinado método contra datasets y benchmarks conocidos, como por ejemplo EVALL. Tratando de ir un poco más allá, están surgiendo iniciativas que proporcionan plataformas de Evaluación como Servicio (EaaS), como por ejemplo TIRA, donde todos los investigadores comparten un entorno de experimentación equivalente, lo que redunda en la mejora de la comparabilidad de sus resultados, y por ende, de la reproducibilidad de sus experimentos. Y esta preocupación por la reproducibilidad está llevando a que cada vez más foros, conferencias y laboratorios promuevan la compartición del código fuente utilizado para realizar los experimentos. Para ello, se dispone de diversas herramientas donde poder compartir el código como GitHub o CRAN, entre otras.

7 REFERENCIAS

[1] Nicola Ferro, Maria Maistro, Tetsuya Sakai, Ian Soboroff. CENTRE@CLEF2018: Overview of the Replicability Task. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. Linda Cappellato, Nicola Ferro, Jian-Yun Nie, Laure Soulier (eds.), CEUR-WS.org, Vol-2125, Avignon, France, September 10-14, 2018.

[2] Enrique Amigó, Jorge Carrillo-de-Albornoz, Julio Gonzalo, Felisa Verdejo. A framework for information systems evaluation, Procesamiento del Lenguaje Natural, Revista nº 57, pp. 189-192, 2016.

[3] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.1:

[10] https://sourceforge.net
[11] https://cran.r-project.org
[12] https://github.com/pan-webis-de

Servicio online. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[4] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.2: Diseño Técnico de la Adaptación de EvALL. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[5] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.3: Código Fuente. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[6] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.4: Informe con los resultados de la campaña IberEval. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[7] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.6: Lecciones aprendidas, futuras mejoras y ampliaciones. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[8] Jorge Carrillo-de-Albornoz y Julio Gonzalo (2018). Servicio para la Puesta en Marcha de una Infraestructura de Evaluaciones Competitivas para Tecnologías del Lenguaje. Entregrable ET.7: Resumen de la adaptación de EvALL. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[9] Paolo Rosso y Francisco Rangel (2018). Implicaciones del Reglamento General de

Protección de Datos en la Organización de Tareas de Evaluación. Plan TL. Plan de Impulso de las Tecnologías del Lenguaje. Secretaría de Estado para el Avance Digital. Ministerio de Economía y Empresa. Gobierno de España.

[10] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA Integrated Research Architecture. Chapter in book on CLEF@20, Springer-Verlag (in press).

8 GLOSARIO DE SIGLAS Y ACRÓNIMOS

CENTRE	ClEf/Ntcir/Trec REproducibility
TIRA	Tool for Integrated Research Architecture